WORKLOAD METRICS FOR SYSTEM EVALUATION

F. Thomas Eggemeier

Wright State University

and

Systems Research Laboratories, Inc.
Dayton, Ohio

INTRODUCTION

A major function of human factors engineering throughout the system development process is to ensure that system demands do not exceed the information processing capabilities of the human operator. Processing overload is a central factor leading to breakdowns in operator performance and to the compromises in system safety and effectiveness that can result from such decrements. Mental workload is the term which refers to that portion of an operator's limited processing capacity which is actually required to perform a particular task or system function. The principal objective of workload assessment is to specify the amount of expended processing capacity so that existing or potential overloads can be identified and decrements in operator performance avoided.

The use of advanced display and control technologies in modern weapons systems has been accompanied in many instances by substantial increases in the monitoring, supervisory, and decision-making demands imposed on the operator. These heavy demands have markedly increased the likelihood of approaching or actually exceeding operator processing capacity limits. As a consequence, assessment of the mental workload imposed by alternative design options has become particularly critical throughout the weapon system design process.

Because of its critical role in the system development process, workload assessment has been the subject of considerable research over the past 10 years (e.g., Moray, 1979). One product of these research efforts has been the development and application of a large number of individual workload assessment techniques. A recent comprehensive review (Wierwille and Williges, 1978) of the workload assessment literature, for example, identified 28 different techniques that had been used to derive measures of load. A substantial number of these empirical assessment techniques can be classified as belonging to one of three categories of workload measures: (1) subjective opinion procedures, (2) performance-based techniques, and (3) physiological techniques.

Subjective techniques (e.g., Gartner and Murphy, 1976; Williges and Wierwille, 1979; Moray, 1982) require that the operator judge and report the degree of workload experienced during performance of a particular task or system

function. Rating scales are the most frequently used type of subjective measurement technique.

Performance-based techniques (e.g., Gartner and Murphy, 1976; Williges and Wierwille, 1979) use some measure of operator behavior or activity as the basis of a workload index. A number of individual assessment techniques can be categorized as performance-based measures. So-called primary task techniques (e.g., Rolfe, 1976; Gartner and Murphy, 1976; Williges and Wierwille, 1979) examine some aspect of the operator's capability to perform the task or system function of interest in order to provide an estimate of load. Deviations from glideslope by a pilot on final approach would constitute one such primary task measure. A second type of performance-based measure which has been frequently used to assess workload is secondary task methodology (e.g., Knowles, 1963; Rolfe, 1971; Ogden, Levine, and Eisner, 1979; Williges and Wierwille, 1979). This approach derives an estimate of workload from the operator's capability to perform a secondary task concurrently with the primary task of interest.

Physiological techniques (e.g., O'Donnell, 1979; Wierwille, 1979) measure some aspect of the operator's physiological response to task or system demand, and provide a measure of load based on these responses. A wide variety of physiological measures (e.g., heart rate variability, pupil diameter, event-related brain potentials) have been used in order to assess workload.

Since a variety of workload assessment procedures are available, an important decision faced by a system designer involves choice of the technique that best meets design requirements. The system development process typically involves a series of stages which range from conceptual development through operational test and evaluation of the system. These stages can be characterized by variations in both the specific questions addressed by workload measurement, and in the practical constraints that must be satisfied by assessment techniques. These questions and constraints suggest a number of criteria that should be considered in choosing a workload measure for application during system development. The purposes of this paper are to outline a set of such criteria, briefly review the current status of the three classes of empirical techniques as they relate to the proposed criteria, and suggest some applications for each class of technique during system development. Some recent work with a subjective assessment procedure which has the potential for application throughout the system development process is also discussed.

WORKLOAD METRIC SELECTION CRITERIA

A number of criteria for evaluation of workload metrics have been proposed in the recent literature (e.g., Gartner and Murphy, 1976; Rolfe, 1976; Ogden et al., 1979; Williges and Wierwille, 1979; Wickens, 1981; Shingledecker, 1983). Several of the proposed criteria are particularly relevant for choice of a metric during system design. These criteria include: (1) sensitivity, (2) diagnosticity, (3) intrusiveness, (4) implementationn requirements, and (5) operator acceptance.

Sensitivity

Sensitivity refers to the capability of a measure to distinguish different levels of load imposed by a task or design option. The degree of sensitivity required in an assessment technique is directly related to the nature of the question to be answered by the workload measure. There are a wide variety of specific design questions (e.g., adequacy of control/display design, allocation of

functions between operators) that can be addressed by workload assessment during system development. Regardless of the specific aspect of the design that is addressed, however, the two basic objectives of workload assessment are to determine: (1) if an overload that would lead to degraded operator performance actually exists, or (2) if the potential for such an overload exists. Questions involving the first objective can be addressed through primary task performance measures, since they are generally assumed to differentiate overload from nonoverload situations (e.g., Knowles, 1963; Gartner and Murphy, 1976; Williges and Wierwille, 1979). In other applications, however, a designer might wish to evaluate the potential for overload among several design options that yield adequate operator performance. This objective is relevant when it is anticipated that other factors during system operation (e.g., environmental stressors, equipment failures) might contribute additional load that would be sufficient to cause degraded operator performance. In this instance, even though none of the design options themselves overload the operator, it is desirable to identify the option that imposes the lowest load and affords the greatest reserve capacity for dealing with other sources of demand. This type of evaluation would require a workload measure that was more sensitive to variations in load than primary task measures, and would suggest the use of other procedures (e.g., subjective, physiological, secondary task) that are designed to discriminate levels of workload in nonoverload situations. Current evidence indicates, for example, that both secondary task measures (e.g., Schifflet, Linton, and Spicuzza, 1982) and subjective ratings of load (e.g., Eggemeier, Crabtree, and LaPointe, 1983) can discriminate differences in task demand that are not reflected in primary task measures of operator performance. The sensitivity criterion is, therefore, an essential consideration in choice of a workload measure, since the degree of sensitivity bears directly on the type of question that can be addressed by a technique.

Diagnosticity

Diagnosticity (Wickens, 1981; Wickens and Derrick, 1981; Shingledecker, 1983) is a second important consideration in choice of a system evaluation metric. This criterion is based on the multiple resources theory (e.g., Navon and Gopher, 1979; Sanders, 1979; Wickens, 1981) explanation of limitations within the human processing system. Essentially, this theory holds that the processing capacity expended in task performance is not unitary, but is drawn from multiple sources or pools, each with its own resources that cannot be exchanged with other pools. One version of multiple resources theory (Wickens, 1981) maintains that perceptual and central processing stages within the human system draw on one resource pool, while the response or motor output stage draws from a separate resource pool. Under this position, it is possible to overload or fully expend the resources associated with one source, while not depleting the processing resources of another source. For example, the requirement to monitor a display which places heavy demands on short-term memory might overload perceptual/central processing resources, while making minimal demands on motor output resources. Other system requirements such as a final approach in an aircraft would have a different demand composition, and might require greater expenditures of motor output resources. Diagnosticity refers to the capability of a technique to discriminate these differences in the load imposed on specific operator resources.

It has been proposed (Wickens, 1981; Wickens and Derrick, 1981) that workload measures vary in their degree of diagnosticity. There are data which indicate, for example, that some physiological measures such as pupil diameter (e.g., Beatty and Kahneman, 1966; Jiang and Beatty, 1981) and some subjective rating scales (e.g., Reid, Shingledecker, and Eggemeier, 1981a; Eggemeier, Crabtree, Zingg, Reid, and Shingledecker, 1982; Notestine, 1983; Wierwille and Casali, 1983a) are

sensitive to perceptual, central processing, and response load manipulations. The event-related brain potential (Isreal, Chesney, Wickens, and Donchin, 1980; Isreal, Wickens, Chesney, Donchin, 1980) and some secondary tasks (e.g., North, 1977; Wickens and Kessel, 1980; Shingledecker, Acton, and Crabtree, 1983), however, show differential sensitivity to manipulations of perceptual/central processing and motor output demands. These data imply that subjective rating scales and some physiological measures are not particularly diagnostic, and can prove sensitive to variations in resource expenditure anywhere within the human processing system. However, other physiological metrics and various secondary tasks appear to be more diagnostic of specific types of resource or capacity expenditure.

Such differences in diagnosticity suggest that the different types of measures can play complementary roles during system development. Less diagnostic measures could serve as screening devices to initially determine if high levels of loading exist during performance of a task or system function, while more diagnostic procedures could be subsequently used to pinpoint the particular source (e.g., perceptual versus motor output) of any such overloads. Choice of an assessment technique on the basis of the diagnosticity criterion would, therefore, be dependent on the objective to be met by the measure of workload.

Intrusiveness

While the criteria of sensitivity and diagnosticity relate to the nature of the question that is to be addressed by a workload measure, there are a number of additional criteria that are suggested by practical constraints imposed on the use of metrics during the system development process. The characteristic of intrusiveness (e.g., Gartner and Murphy, 1976; Williges and Wierwille, 1979; Shingledecker, 1983) is one such criterion, and refers to the tendency for some metrics to cause degradations in ongoing primary task performance.

Intrusiveness in an assessment procedure is undersirable on both practical and theoretical grounds. From a practical perspective, it is clear that any technique that causes decrements in operator performance can potentially compromise the safety of system operation. Such compromises are obviously unacceptable, particularly during the later stages of system development when operational test and evaluations of prototype or initial production models are conducted. From a theoretical point of view, intrusiveness can cause problems in the interpretation of data resulting from application of an assessment technique. These interpretation problems stem from the assumption that measurement procedures provide a pure index of the load imposed by the primary task. If primary task performance is degraded by the introduction of the assessment technique, an unbiased measure of primary task workload is not possible. Although intrusiveness presents potential difficulties for all metrics, the interpretation problem can be particularly acute with secondary task measures (Rolfe, 1971; Ogden et al., 1979) that are intended to provide a measure of the reserve capacity afforded by the primary task.

Despite its importance, the comparative data base on the degree of intrusion associated with the various types of metrics is not extensive. Some significant steps toward establishing a systematic data base have been undertaken recently (e.g., Casali and Wierwille, 1982, 1983; Rahimi and Wierwille, 1982; Shingledecker, Crabtree, and Acton, 1982; Acton, Crabtree, and Shingledecker, 1983; Wierwille and Casali, 1983b; Wierwille and Conner, 1983), but such direct comparison data are not yet complete. However, some statements regarding the potential for intrusiveness can be made on the basis of data generated by individual applications of the various techniques.

First, it is clear that intrusiveness has represented a major problem in many applications of secondary task methodology (e.g., Rolfe, 1971; Gartner and Murphy, 1976; Ogden et al., 1979; Williges and Wierwille, 1979). The problem has led to the development of techniques such as cross-adaptive (e.g., Kelly and Wargo, 1967; Jex and Clement, 1979) and embedded (Shingledecker, Crabtree, Simons, Courtright, and O'Donnell, 1980; Shingledecker, 1980; Crabtree and Spicuzza, 1981; Shingledecker and Crabtree, 1982) secondary tasks that are designed to minimize or control the levels of intrusion. Cross-adaptive procedures permit variations in secondary task difficulty as a function of primary task performance. When primary task performance falls below a specified criterion, secondary task difficulty is reduced in order to control the level of intrusion. This type of procedure has been successfully employed in a number of laboratory and simulation studies (Kelly and Wargo, 1967; Jex and Clement, 1979) that have utilized primary continuous tracking tasks. Applications of the procedure to discrete tasks in more complex environments have not been accomplished, and could present difficulties due to problems in obtaining primary task measures that would permit adaptation of the secondary task. The embedded secondary task approach, on the other hand, was developed for application to high fidelity simulation or operational environments. This procedure uses an element already embedded in normal system operation procedures as the secondary task. The elements chosen as secondary tasks (e.g., radio communications) are those that are normally assigned lower priority than the primary task (e.g., flight control), thereby minimizing the potential for primary task intrusion.

Second, it appears that the intrusion associated with most other classes of assessment techniques tends to be minimal. Subjective assessment techniques typically present no significant intrusion problem, since rating scales and other report procedures are usually completed subsequent to primary task performance. Primary task measures are, by definition, nonintrusive, because their application involves no additional operator performance or reports. Physiological procedures also appear to minimize the potential for intrusion, although there are data (Rahimi and Wierwille, 1982) which indicate that these techniques can be associated with some intrusion.

The degree of intrusiveness that can be tolerated in an assessment technique will vary as a function of the context in which the measure is taken. Some degree of intrusion in a simulator or in a crewstation mockup could be less serious, for example, than equivalent levels of primary task decrement during actual system operation. Choice of an assessment procedure on the basis of intrusiveness would, therefore, be determined in part by constraints dictated by the measurement situation.

Implementation Requirements

The implementation requirements associated with a particular measurement technique constitute a second criterion that is heavily influenced by the practical constraints imposed by the system development process. Implementation requirements are factors that are related to the ease with which a technique can be applied at different stages of system development and evaluation. Examples of such factors include: (1) the instrumentation and software that is required to record and analyze the measures associated with a technique; (2) any operator training that is necessary for the technique to be properly applied; and (3) system simulation facilities or actual equipment that are required for application of the technique.

Different classes of assessment procedures can vary considerably in their instrumentation requirements, as can individual techniques within the same category. For instance, subjective opinion measures usually make use of paper and pencil for data recording, while much more stringent implementation requirements are typically associated with physiological and some performance-based procedures. Requirements also vary within categories themselves. Cross-adaptive secondary techniques require more extensive instrumentation than other secondary task procedures (e.g., interval production, Shingledecker et al., 1983) which require only a means of recording an operator's response. Therefore, when minimal instrumentation is a primary constraint, the use of subjective measures or certain secondary task procedures such as the interval production task is suggested.

Operator training requirements also vary with techniques and can be necessary with both secondary task and subjective assessment procedures. Applications of secondary task methodology, for instance, usually require some operator training in order to stabilize baseline performance on the secondary task before it is performed concurrently with the primary task. Some subjective procedures (e.g., Reid et al., 1981a) also include the provision for familiarization with the rating scales prior to their use. Training requirements associated with the use of both primary task and physiological measures would be virtually nonexistent in most cases.

Techniques can also differ in the types of simulation facilities and operational equipment that are necessary for their application. Such facility requirements can be particularly restrictive during the early conceptual stage of system development, when system design information is very general, and simulation and mockup facilities are typically not available. Since both performance-based and physiological techniques require such facilities, their application has been usually restricted to later stages (e.g., validation, engineering development) of the design process when the appropriate devices are present. This constraint on early use of physiological and performance-based procedures is one factor that has led to the development and application of analytical time-line techniques (e.g., Zipoy, Premselaar, Gargett, Belyea, and Hall, 1970; Parks, 1979; Geer, 1981) and several simulation models (e.g., Linton, Jahns, and Chatelier, 1977; Lane, Strieb, and Wherry, 1977; Lane, Strieb, Glenn, and Wherry, 1981; Chubb, 1981) that are capable of addressing workload assessment issues during earlier stages of design. Traditional applications of subjective metrics also require the availability of mockups, simulators, or operational equipment. However, a recent application (Quinn, Jauer, and Summers, 1982) demonstrated the projective use of a subjective metric by requiring experienced pilots to rate the expected load associated with several proposed cockpit enhancements. The projective ratings were based on detailed descriptions of mission profiles and control/display options, and were intended to provide workload estimates that could be combined with other factors (e.g., cost) to initially screen design options for further evaluation. Although the results must be validated, the Quinn et al. study provides a methodology with the potential to permit application of subjective procedures during the earlier stages of development when performance-based and physiological techniques are not practicable.

Taken together, implementation requirements can therefore impose important constraints on the use of the various classes of assessment techniques during the development process. Instrumentation and facility requirements are typically more stringent with performance-based and physiological techniques than with subjective procedures, suggesting the use of the latter for certain situations.

Operator Acceptance

The characteristic of operator acceptance is important to ensure that an assessment technique will yield data that are representative of the load imposed by the task or system function in question. Assessment procedures which are perceived by operators as bothersome or artificial incur the risk of being ignored, performed at substandard levels, or being associated with significant levels of primary task intrusion. Any of these factors can lead to compromises in the effectiveness of a technique.

In spite of the potential importance of operator acceptance, there are little or no formal comparative data which are available to address operator reaction to the major classes of techniques. Although some investigators (e.g., Hallsten and Borg, 1975) have commented on operator acceptance of a number of procedures, the data are not sufficient to address the issue in a comprehensive manner. Informal data and knowledge of the procedures involved in application of the techniques can, however, be used to provide some estimates of acceptance. Informal evidence, for example, suggests that subjective procedures usually enjoy a high degree of user acceptance, quite possibly because of the high face validity associated with many current rating scales (e.g., Cooper and Harper, 1969; Reid et al., 1981a). Operator acceptance should also be quite good for primary task measures, since they do not typically involve any additional operator response or effort. Physiological techniques would have some potential for low acceptance if the recording instruments used are considered bothersome by the operator, but this does not appear to have been a significant problem with most techniques. Secondary task methods could also be considered distracting by the operator if the requirement to perform the secondary task interferes with primary task performance. The embedded secondary task technique (Shingledecker et al., 1980) which utilizes a secondary task that is normally performed in the operational environment should, however, minimize this risk.

APPLICATION GUIDELINES

It is obvious from the foregoing discussion that no single assessment technique is capable of meeting all of the criteria outlined above. The various categories of techniques are characterized by the capability to satisfy some criteria, but not others. Criteria vary in their importance as a function of the different stages of design, and consequently, techniques vary in their applicability. It is therefore clear that assessment of workload across the various phases of the design process will require the complementary use of multiple metrics, since no single metric is capable of providing all of the required information.

The capability of individual assessment procedures to meet the various criteria can provide some guidance regarding their use for specific purposes at different stages of design. Table 1 summarizes the current status of the procedures with respect to the proposed criteria, and can be used as a basis to suggest particular applications for each class of technique.

An investigator requiring a nonintrusive general measure of load in an operational environment with restricted data recording capabilities should, for example, consider the application of subjective metrics. On the other hand, primary task measures might be considered for application in a high fidelity simulator with performance measurement capability when the objective was to evaluate the adequacy of operator performance with a particular design option. The use of secondary task methodology or an appropriate physiological technique in a system simulator would be suggested if the intent was to isolate the source of

TABLE 1. SUMMARY OF WORKLOAD ASSESSMENT TECHNIQUE CAPABILITIES

	Sensitivity	Diagnosticity	Intrusiveness	Implementation Requirements	Operator Acceptance
PRIMARY TASK MEASURES	Discriminate overload from nonoverload situations. Used to determine if operator performance will be acceptable with a particular design option.	Not considered diagnostic. Represents a global measure of workload that is sensitive to overloads anywhere within the operator's processing system.	Nonintrusive since no additional operator performance or report required.	Instrumentation for data collection can restrict use in operational environments. Use requires mockups, simulators, or operational equipment. Imposes limits on use during early system development. No operator training required.	No systematic data. No reason to expect negative operator opinion.
SECONDARY TASK METHODS	Capable of discriminating levels of capacity expenditure in nonoverload situations. Used to assess reserve capacity afforded by a primary task. Can be used to assess the potential for overload among design options.	Capable of discriminating some differences in resource expenditure (e.g., central processing versus motor). Diagnosticity suggests complementary use with more generally sensitive measures, with the latter initially identifying overloads and secondary tasks being used subsequently to pinpoint the locus of overload.	Primary task intrusion has represented a problem in many applications, particularly in the laboratory. Data are not extensive in operational environments. Several techniques (e.g., embedded secondary task, adaptive procedures) have been designed to control intrusion. Potential for intrusion could limit use in operational environments.	Instrumentation for data collection can restrict use in operational environments, but some tasks have been instrumented for in-flight use. Use requires mockups, simulators, or operational equipment. Imposes limits on use during early system development. Some operator training usually required to stabilize secondary task performance.	No systematic data. Requirement to perform secondary task could distract operator. Technique such as embedded secondary task should minimize any acceptance problems.
PHYSIOLOGICAL TECHNIQUES	Capable of discriminating levels of capacity expenditure in nonoverload situations. Can be used to assess the relative potential for overload among design options.	Some techniques (e.g., event-related brain potential) appear diagnostic of some resources, while other measures (e.g., pupil diameter) appear more generally sensitive. Choice of technique dependent on purpose of measurement (screening for any overload versus identifying locus of overload).	Intrusion does not appear to represent a major problem, although there are data to indicate that some interference can occur.	Instrumentation for data collection can restrict use in operational environments. Use requires mockups, simulators, or operational equipment. Imposes limits on use during early system development. No operator training required.	No systematic data. Instrumentation and recording equipment could represent potential problem, but no significant problems reported in literature.
SUBJECTIVE TECHNIQUES	Capable of discriminating levels of capacity expenditure in nonoverload situations. Can be used to assess the relative potential for overload among design options.	Not considered diagnostic. Available evidence indicates that rating scales represent a global measure of load. Lack of diagnosticity suggests use as a general screening device to determine if overload exists anywhere within task performance.	Intrusion does not appear to represent a significant problem. Most applications require rating scale completion subsequent to task performance and, therefore, present no intrusion problem.	Instrumentation required is usually minimal, permitting use in a number of environments. Traditional applications require mockups, simulators, or operational equipment. Imposes limits on use during early system development. Recent projective use provides potential for application during early stages. Some familiarization with procedures can be required.	No systematic data. Informal evidence suggests that several rating scales enjoy a high degree of operator acceptance.

an overload that had been previously identified through use of a subjective or primary task metric. The potential applications for each class of metric are clearly much more extensive than those suggested by these hypothetical situations. The examples do, however, illustrate how the proposed criteria can be applied to identify the class(es) of techniques that might be most appropriate for a particular application.

In many instances, use of the proposed criteria will result in simultaneous application of more than one technique. Applications of secondary task methodology, for example, require the measurement of primary task performance in order to evaluate the degree of any intrusion that might have occurred. In other instances, the objectives of an evaluation might also suggest the concurrent use of more than one metric. For example, a comprehensive evaluation of two display options might include the use of both primary task and subjective measures. The primary task measure would permit assessment of any differences in the adequacy of task performance that could be expected with the options, while the subjective technique would provide the potential to identify any workload differences between the options that were not reflected in the less sensitive performance measure.

The preceding review and discussion of metrics has been primarily concerned with classes of workload assessment techniques in general. It is clear from the foregoing discussion, for example, that the general category of subjective metrics holds a great deal of potential for use during system design. Once a class of technique has been identified as appropriate, however, an individual procedure or measure from within the category must be chosen for actual application. Individual procedures themselves can also vary along a number of dimensions that can impact their suitability for use. The purpose of the following discussion is to briefly review some recent work with an individual subjective assessment technique that appears to be particularly well suited for a number of applications throughout the system development process.

SUBJECTIVE WORKLOAD ASSESSMENT

Subjective workload measurement procedures satisfy a number of the criteria outlined above and, as a consequence, have been very frequently employed as workload assessment techniques (e.g., Williges and Wierwille, 1979). Despite their advantages, there are several problems which have been traditionally associated with use of subjective workload metrics. First, in many applications, individual rating scales have been developed for a specific investigation and have not been validated for generalized use. Second, there is little evidence in the literature of workload rating scales that have been rigorously developed on the basis of psychometric procedures (e.g., Williges and Wierwille, 1979). As a consequence, most available rating scales have unknown metric properties, and must be assumed to provide only ordinal level measurement.

In order to provide a workload rating scale with known metric properties and with the potential for generalized applicability, a procedure termed the Subjective Workload Assessment Technique (SWAT) has been developed (Reid et al., 1981a; Reid, Shingledecker, Nygren, and Eggemeier, 1981b; Reid, Eggemeier, and Shingledecker, 1982). In SWAT, it is assumed that there are three major contributors to subjective mental load: (1) time load, (2) mental effort load, and (3) psychological stress load. Time load refers to the percentage of time that an operator is busy, and reflects such factors as overlap and interruption among tasks. Mental effort load, on the other hand, refers to the degree of attention or concentration required during task performance. The final dimension, psychological stress load, reflects any additional factors that cause operator anxiety

or confusion and, therefore, contribute to subjective mental load. In SWAT, each of the three dimensions is represented by an individual three-point rating scale with verbal descriptors that define the levels on each dimension.

SWAT is based on application of conjoint measurement and scaling (e.g., Nygren, 1982). Conjoint measurement and scaling permit ratings on the three dimensions to be combined into one overall scale of workload with interval measurement properties. In order to identify the rule which is appropriate for combining the three dimensions into the overall interval scale, a scale development phase is completed. During this phase, subjects rank-order the subjective load associated with the 27 possible combinations that result from the three levels of time, mental effort, and psychological stress load. This rank-ordering information is subjected to a series of axiom tests to identify the rule for combining the three dimensions. When the rule has been established, conjoint scaling is applied to derive the overall scale of workload. Subsequent to the scale development phase, subjects participate in an event scoring phase. During event scoring, subjects perform the task or mission segment of interest and rate the time, mental effort, and stress load associated with performance. The ratings on the individual dimensions are then converted to one of the 27 points on the interval scale that was derived during scale development. More extensive discussions of the scale development and event scoring procedures can be found in Reid et al. (1981a,b), and Reid, Eggemeier, and Nygren (1982).

One aspect of the work conducted during the development of SWAT has centered on establishing its capability to reflect workload differences in a number of different types of tasks in several environments that are representative of those found during system development. SWAT has been successfully applied in a number of laboratory or part-task simulation environments (e.g., Reid et al., 1981a; Eggemeier et al., 1982, 1983; Notestine, 1983); in several full mission simulators (e.g., Reid, Eggemeier, and Shingledecker, 1984; Skelly, Reid, and Wilson, 1983); and under conditions that are similar to the early stages of system development when workload estimates must be based on detailed mission scenarios and descriptions of system equipment capabilities (Quinn et al., 1982).

Figure 1 illustrates the results of two applications of SWAT in laboratory/part-task simulation environments. Panel A (Reid et al., 1981a) shows the results of an experiment which employed several levels of a simulated flight control (critical tracking, Jex and Clement, 1979) task and a secondary simulated aircrew radio communications task (Shingledecker et al., 1980). Significant differences in SWAT ratings were obtained in the communication task alone condition versus the more difficult dual task condition. SWAT ratings also successfully discriminated levels of difficulty in both the simulated flight control and radio communications tasks. Panel B (Eggemeier et al., 1983) illustrates the effects on SWAT ratings of variations in the rate of stimulus presentation in a sequential short-term memory task. Subjects in the experiment were required to monitor a visual display and update the status of four categories of information that changed at several rates. The memory task was intended to be representative of the demands placed on air traffic controllers while monitoring flight control displays. SWAT ratings successfully discriminated levels of difficulty in the memory task, even though a primary task measure of performance errors showed no significant differences between conditions.

Several recent experiments also support the applicability of SWAT to full mission simulation environments. SWAT ratings have proven sensitive to expected workload variations in high fidelity flight simulation evaluations of advanced control/display options in both fighter (Reid et al., 1984) and bomber (Skelly

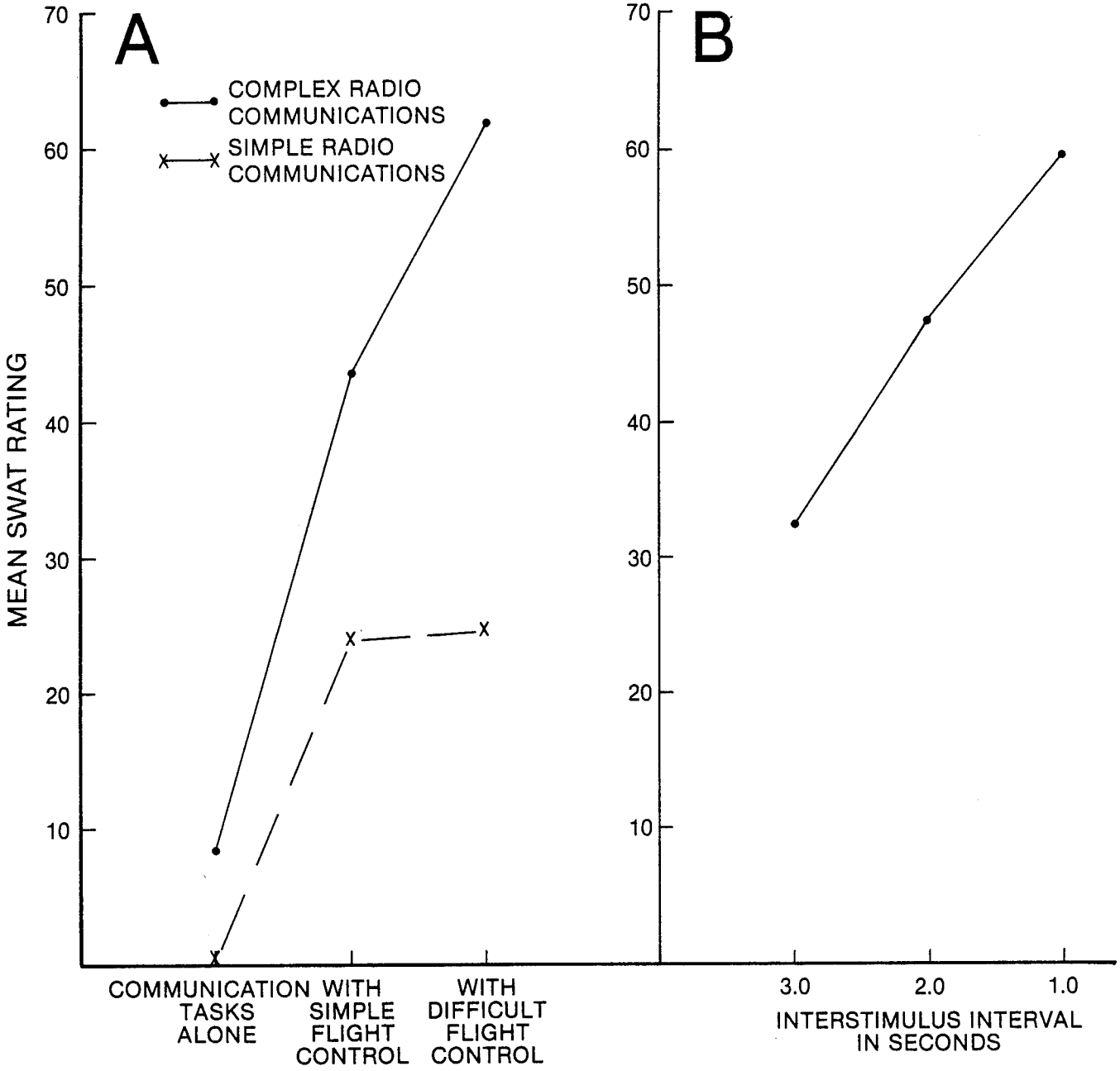

Figure 1. Mean SWAT Ratings as a Function of Task Difficulty in Two Experiments. [Panel A illustrates the effects of simple and complex radio communications on SWAT ratings in both single and dual task conditions (Figure drawn from the data of Reid et al., 1981a). Panel B shows the effect of stimulus presentation rate manipulations in a sequential short-term memory task (Figure adapted from Eggemeier et al., 1983).]

et al., 1984) aircraft. Reid et al., for example, obtained significant differences in pilot SWAT ratings as a function of variations in the number of opponents during a fighter mission. SWAT ratings in the Skelly et al. study also showed differences that were logically defensible and consistent with expectations. Pilot ratings, for instance, were generally higher than copilot ratings, except for a number of segments in which the copilot was flying the aircraft. Segments of simulated mission which included various types of threats to the aircraft were rated higher than baseline segments that did not include such threats. In both applications, pilot acceptance of the rating procedure was very high, and in both instances, SWAT ratings were taken with minimal intrusion by having the pilot verbally report ratings after completion of a mission segment to an experimenter stationed outside the cockpit.

The Quinn et al. (1982) experiment that was briefly discussed earlier also utilized the SWAT methodology in a novel application of the technique. The purpose of the Quinn et al. study was to evaluate a variety of methods for enhancing fighter aircraft systems, including advanced display, control, and navigational concepts. A methodology was devised to comparatively evaluate the enhancements along a number of dimensions prior to prototype development, and SWAT was included to quantify predicted effects on pilot workload. A number of experienced fighter pilots were provided with a mission scenario and detailed descriptions of an advanced baseline version of the aircraft and several enhancements. On the basis of the information, the pilots provided mission SWAT ratings for the various versions of the baseline system that included several combinations of enhancements. The interval level data that are obtained from the SWAT procedure permitted use of the resulting workload ratings in a multiattribute utility analysis with other factors (cost, system performance) to permit selection of several options for further research. Although it is clear that the results of the projective SWAT ratings must be validated, the methodology employed by Quinn et al. is significant in that it demonstrates the feasibility of obtaining SWAT ratings on the basis of detailed mission and equipment information. Use of the technique in this manner includes obvious time and cost advantages and, as noted previously, demonstrates the potential for application of SWAT during the earlier stages of system design.

Taken together, the results of current work with the SWAT technique clearly support its sensitivity to a variety of tasks that are relevant to system operation. The available evidence also indicates that SWAT has a very high potential for applicability across several stages of design. These data, coupled with the advantages of the interval level measurement afforded by the technique, strongly support the utility of the SWAT metric for evaluation of workload during the system development process.

SUMMARY AND CONCLUSIONS

Application of the proposed criteria to the major categories of workload assessment techniques indicates that a battery of performance-based, subjective, and physiological metrics will be required to meet the varied needs for workload measurement that arise during the system development process. In many instances, the capabilities of one technique supplement those of another procedure, suggesting the complementary use of the various metrics at different stages of design. Among the classes of assessment procedures reviewed above, subjective techniques appear to have the greatest potential for application across the various phases of the design process, and the SWAT technique is one such procedure that has demonstrated high levels of sensitivity and applicability.

Although current information is sufficient to suggest some applications for the various categories of techniques, more extensive data are needed to refine procedures for choice of a metric for particular applications. For example, more complete comparative data on the relative sensitivity and intrusiveness that can be expected from individual techniques from within particular categories (e.g., secondary task) of procedures represent a need in this area. As was noted previously, several such efforts have been recently undertaken (e.g., Shingledecker et al., 1983; Wierwille and Casali, 1983b), and the results should provide a more refined basis for choice of metric for particular applications. An additional area requiring further experimentation deals with the extension of current classes of metrics to the earlier and later stages of the design process. Implementation requirements have somewhat limited the applicability of secondary task and physiological metrics in the early and latter stages of system design, and more work is required to evaluate the application of these techniques beyond the laboratory and simulation environments. Some of this type of work (e.g., Schifflet et al., 1982) has been conducted, but additional efforts are required. Further evaluation and extension of the Quinn et al. (1982) procedure for application of subjective metrics during the early stages of design should also be pursued in order to supplement available analytic and modeling procedures that provide the current capability for workload assessment during this phase of the development process.

Release

This paper has been approved for public release, distribution unlimited.

Acknowledgement

This work was partially supported by the USAF Aerospace Medical Research Laboratory under Contract No. F33615-82-K-0512. Mr. Mark S. Crabtree provided very helpful comments on an earlier version of this paper.

REFERENCES

Acton, W. H., Crabtree, M. S., and Shingledecker, C. A., 1983, Development of a standardized workload metric evaluation methodology, Proceedings of the 1983 IEEE National Aerospace and Electronics Conference, 1086-1089.

Beatty, J. and Kahneman, D., 1966, Pupillary changes in two memory tasks, Psychonomic Science, 55:371-372.

Casali, J. G. and Wierwille, W. W., 1982, A sensitivity/intrusion comparison of mental workload estimation techniques using a flight task emphasizing perceptual piloting activities, Proceedings of the 1982 IEEE International Conference on Cybernetics and Society, 598-602.

Casali, J. G. and Wierwille, W. W., 1983, Communications-imposed pilot workload: A comparison of sixteen estimation techniques, Proceedings of Second Ohio State University Symposium on Aviation, 223-235.

Chubb, G. P., 1981, SAINT, A digital simulation language for the study of manned systems: in: "Manned Systems Design Methods, Equipment, and Applications," J. Morssel and K. F. Kraiss, eds., Plenum Press, New York.

Cooper, G. E. and Harper, R. P., Jr., 1969, The use of pilot rating in the evaluation of aircraft handling qualities, Report No. NASA TN-D-5513, National Aeronautics and Space Administration, Ames Research Center, Moffett Field, California.

Crabtree, M. S. and Spicuzza, R. J., 1981, Evaluation of embedded radio communications activities as secondary tasks for objective assessment of aircrew workload in simulators, trainers, and actual systems. Proceedings of the 1981 IEEE National Aerospace and Electronics Conference, 1349-1352.

Eggemeier, F. T., Crabtree, M. S., Zingg, J. J., Reid, G. B., and Shingledecker, C. A., 1982, Subjective workload assessment in a memory update task, Proceedings of the 1982 Human Factors Society Annual Meeting, 643-674.

Eggemeier, F. T., Crabtree, M. S., and LaPointe, P. A., 1983, The effect of delayed report on subjective workload ratings, Proceedings of the 1983 Human Factors Society Annual Meeting, 139-143.

Gartner, W. P. and Murphy, M. R., 1976, Pilot workload and fatigue: A critical survey of concepts and assessment techniques, Report No. NASA-TN-D-8365, National Aeronautics and Space Administration, Washington, D.C.

Geer, C. W., 1981, Human engineering procedures guide, Technical Report No. AFAMRL-TR-81-35, USAF Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio.

Hallsten, L. and Borg, G., 1975, Six rating scales for perceived difficulty, Report #58 from the Institute of Applied Psychology, The University of Stockholm, Stockholm, Sweden.

Isreal, J. B., Chesney, G. L., Wickens, C. D., and Donchin, E., 1980, P300 and tracking difficulty: Evidence for multiple resources in dual-task performance, Psychophysiology, 17:259-273.

Isreal, J. B., Wickens, C. D., Chesney, G. L., and Donchin, E., 1980, The event-related brain potential as an index of display-monitoring workload, Human Factors, 22:211-244.

Jex, H. R. and Clement, W. F., 1979, Defining and measuring perceptual-motor workload in manual control tasks: in: "Mental Workload: Its Theory and Measurement," N. Moray, ed., Plenum Press, New York.

Jiang, Q. and Beatty, J., 1981, Physiological assessment of operator workload during manual tracking, Proceedings of the 17th Annual Conference on Manual Control, Los Angeles, California.

Kelly, C. R. and Wargo, M. J., 1967, Cross-adaptive operator loading tasks, Human Factors, 9:395-404.

Knowles, W. B., 1963, Operator loading tasks, Human Factors, 5:151-161.

Lane, N. E., Streib, M. I., Glenn, F. A., and Wherry, R. J., 1981, The human operator simulator: An overview: in: "Manned Systems Design Methods, Equipment, and Applications," J. Morssel and K. F. Kraiss, eds., Plenum Press, New York.

Lane, N. E., Streib, M., and Wherry, R. J., Jr., 1977, The human operator simulator: Estimation of workload reserve using a simulated secondary task, Proceedings of the AGARD Conference on Methods to Assess Workload, No. AGARD-CP-216.

Linton, P. M., Jahns, D. W., and Chatelier, P. R., 1977, Operator workload assessment model: An evaluation of a VF/VA-V/STOL system, Proceedings of the AGARD Conference on Methods to Assess Workload, No. AGARD-CP-216.

Moray, N., ed., 1979, "Mental Workload: Its Theory and Measurement," Plenum Press, New York.

Moray, N., 1982, Subjective mental workload, Human Factors, 24:25-40.

Navon, D. and Gopher, D., 1977, On the economy of the human processing system, Psychological Review, 86:214-255.

North, R. A., 1977, Task components and demands as factors in dual-task performance, Report No. ARL-77-2/AFOSR-77-2, Aviation Research Laboratory, University of Illinois at Urbana-Champaign, Champaign, Illinois.

Notestine, J., 1983, Subjective workload assessment in a probability monitoring task and the effect of delayed ratings, Unpublished Master's Thesis, Wright State University, Dayton, Ohio.

Nygren, T. E., 1982, Conjoint measurement and conjoint scaling: A users guide, Technical Report No. AFAMRL-TR-82-22, U.S. Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio.

O'Donnell, R. D., 1979, Contributions of psychophysiological techniques to aircraft design and other operational problems, Report No. 244, NATO Advisory Group for Aerospace Research and Development.

Ogden, G. D., Levine, J. M., and Eisner, E. J., 1979, Measurement of workload by secondary tasks, Human Factors, 21:529-548.

Parks, D. L., 1979, Current workload methods and emerging challenges: in: "Mental Workload: Its Theory and Measurement," N. Moray, ed., Plenum Press, New York.

Quinn, T. J., Jauer, R. A., and Summers, P. I., 1982, Radar aided mission/aircrew capability exploration, RAM/ACE interim report--task II synthesis, Technical Report No. AFAMRL-TR-82-91, U.S. Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio.

Rahimi, M. and Wierwille, W. W., 1982, Evaluation of the sensitivity and intrusion of workload estimation techniques in piloting tasks emphasizing mediational activity, Proceedings of the 1982 IEEE International Conference on Cybernetics and Society, 593-597.

Reid, G. B., Eggemeier, F. T., and Nygren, T. E., 1982, An individual differences approach to SWAT scale development, Proceedings of the 1982 Human Factors Annual Meeting, 639-642.

Reid, G. B., Eggemeier, F. T., and Shingledecker, C. A., 1982, Subjective workload assessment technique, Proceedings of the 1982 AIAA Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics, 281-288.

Reid, G. B., Eggemeier, F. T., and Shingledecker, C. A., 1984, Workload analysis for the AMRAAM operational test and evaluation, Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio (in preparation).

Reid, G. B., Shingledecker, C. A., and Eggemeier, F. T., 1981a, Application of conjoint measurement to workload scale development, Proceedings of the 1981 Human Factors Society Annual Meeting, 522-526.

Reid, G. B., Shingledecker, C. A., Nygren, T. E., and Eggemeier, F. T., 1981b, Development of multidimensional subjective measures of workload, Proceedings of the 1981 IEEE International Confererence on Cybernetics and Society, 403-406.

Rolfe, J. M., 1971, The secondary task as a measure of mental load: in: "Measurement of Man at Work," W. T. Singleton, J. G. Fox, and D. Whitfield, eds., Taylor and Francis, London, 135-148.

Rolfe, J. M., 1976, The measurement of human response in man-vehicle control situations" in: "Monitoring Behavior and Supervisory Control," T. Sheridan and G. Johannsen, eds., Plenum Press, New York.

Sanders, A. F., 1979, Some remarks on mental load: in: "Mental Workload: Its Theory and Measurement," N. Moray, ed., Plenum Press, New York.

Schifflet, S. G., Linton, P. M., and Spicuzza, R. J., 1982, Evaluation of a pilot workload assessment device to test alternative display formats and control handling qualities, Proceedings of the 1982 AIAA Workshops on Flight Testing to Identify Pilot Workload and Pilot Dynamics, 222-233.

Shingledecker, C. A., 1980, Enhancing operator acceptance and noninterference in secondary task measures of workload, Proceedings of the Human Factors Society 1980 Annual Meeting, 674-677.

Shingledecker, C. A., 1983, Behavioral and subjective workload metrics for operational environments, Proceedings of the AGARD (AMP) Symposium, "Sustained Intensive Air Operations: Physiological and Performance Aspects," Paris, France (preprint).

Shingledecker, C. A., Acton, W. H., and Crabtree, M. S., 1983, Development and application of a criterion task set of workload metric evaluation, SAE Technical Paper Series, Paper No. 831419, Society of Automotive Engineers, Warrendale, Pennsylvania.

Shingledecker, C. A. and Crabtree, M. S., 1982, Subsidiary radio communications tasks for workload assessment in R&D circulations: II. Task sensitivity evaluation, Technical Report No. AFAMRL-TR-82-57, U.S. Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio.

Shingledecker, C. A., Crabtree, M. S., and Acton, W. H., 1982, Standardized tests for the evaluation and classification of workload metrics, Proceedings of the 1982 Human Factors Society Annual Meeting, 648-651.

Shingledecker, C. A., Crabtree, M. S., Simons, J. C., Courtright, J. F., and O'Donnell, R. D., 1980, Subsidiary radio communications tasks for workload assessment in R&D simulations: I. Task development and workload scaling, Technical Report No. AFAMRL-TR-80-126, U.S. Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio.

Skelly, J., Reid, G. B., and Wilson, G. R., 1983, B-52 full mission simulation: Subjective and physiological workload applications, Paper presented at the Second Aerospace Behavioral Engineering Technology Conference, Long Beach, California.

Wickens, C. D., 1981, Processing resources in attention, dual task performance, and workload assessment, Technical Report No. EPL-81-3, Engineering Psychology Research Laboratory, University of Illinois, Champaign, Illinois.

Wickens, C. D. and Derrick, W., 1981, Workload measurement and multiple resources, Proceedings of the 1981 IEEE Conference on Cybernetics and Society, 600-603.

Wickens, C. D., and Kessel, C., 1980, The processing resource demands of failure detection in dynamic systems, Journal of Experimental Psychology: Human Perception and Performance, 6:564-577.

Wierwille, W. W., 1979, Physiological measures of aircrew mental workload, Human Factors, 21:575-593.

Wierwille, W. W. and Casali, J. G., 1983(a), A validated rating scale for global mental workload measurement applications, Proceedings of 1983 Human Factors Society Annual Meeting, 129-133.

Wierwille, W. W. and Casali, J. G., 1983(b), The sensitivity and intrusion of mental workload estimation techniques in piloting tasks, IEOR Department Report No. 8309, Department of Industrial Engineering and Operations Research, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Wierwille, W. W. and Connor, S. A., 1983, Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator, Human Factors, 25:1-16.

Wierwille, W. W. and Williges, R. C., 1978, Survey and analysis of operator workload assessment techniques, Report No. S-78-101, Systemetrics, Inc., Blacksburg, Virginia.

Williges, R. C. and Wierwille, W. W., 1979, Behavioral measures of aircrew mental workload, Human Factors, 21:549-574.

Zipoy, D. R., Premselaar, S. J., Gargett, R. E., Belyea, I. L., and Hall, J. J., 1970, Integrated information presentation and control systems study, Vol. I, System development concepts, Air Force Flight Dynamics Laboratory, Technical Report No. AFFDL-TR-70-79, Wright-Patterson Air Force Base, Ohio.